The Metrics Maze: Navigating the Landscape of Evaluation Techniques for MT Systems

Henri UPTON
ENSAE 3A
henri.upton@ensae.fr

Robin GUILLOT
ENSAE 3A
robin.guillot@ensae.fr

https://github.com/henriupton99/nlp-metrics-benchmark

Abstract

Over the years, more and more untrained text similarity metrics have emerged in a context where tasks are becoming increasingly numerous, such as text summarization, story writing, or translation. Compared to trained metrics, they are independent of a training set and must perform well in any context. Therefore, it becomes increasingly important to compare these metrics in order to clearly identify which ones perform the best (Colombo et al., 2022b). In this paper, we focus on the task of Machine Translation (MT) and benchmark the sentence-level correlation of the main existing metrics with human scores, using the WMT22 (World Machine Translation) dataset.

1 Introduction

The automatic evaluation of Machine Translation (MT) has posed a significant challenge in recent years. Over the past decade, there has been a growing recognition of the importance of developing reliable metrics that can accurately assess the quality of MT output. In response to this need, the Workshop on Machine Translation (WMT) was established in 2006, providing an annual forum for researchers to collaborate and share their latest findings on MT evaluation.

One of the key features of the WMT is its competitions, which focus on various aspects of MT, including the development of automatic metrics that can effectively assess the quality of MT output. These competitions require participants to submit their automatic metrics, which are then evaluated against human judgments to determine the best-correlated scores.

Despite the growing interest in automatic evaluation for MT, this remains a challenging issue. The development of reliable metrics that can accurately assess the quality of MT output continues to be an active area of research. The WMT and similar initiatives have played an essential role in advancing the field, providing a platform for researchers to share their latest findings and collaborate on the development of new evaluation techniques.

2 Problem Framing

2.1 Choice of the dataset

We focus our study on the WMT22 database, which provides sets of translations performed by Natural Language Generation (NLG) systems, from a source language (sl) to a target language (tl). We dispose of three pairs of sl-tl languages as illustrated in Figure 1.

For each of these three pairs, we have access to two databases thanks to the Google MQM Human Evaluation GitHub https://github.com/google/wmt-mqm-human-evaluation (Freitag et al., 2021a), which proposes all our variables of interest:

- **"Candidate List"** database: translations candidates from different NLG systems, source sentences, and reference translations established by experts. Note that for each candidate translation, there is only one reference translation. We also have access to the domain and the segment to which belongs each source sentence.
- **"Candidate Correction"** database: lists all errors made by NLG systems; several human experts has annotated for each candidate translation the translation errors, indicating for each error its type and severity, i.e. whether it is a significant error or not.

2.2 Gold references

One of the first criteria to consider when evaluating untrained metrics is their correlation with hu-

SL - TL	# samples	# segments
English-German	32190	2027
English-Russian	32348	2027
Chinese-English	35188	1875

Figure 1: Number of samples, number of segments and average score for each couple source language - target langage

man judgment. The WMT22 data allows us to assign a human score to each candidate translation of the considered systems by calculating the MQM (Multidimensional Quality Metrics) scores which aggregates all errors identified by human experts:

$$MQM_{hyp} = - \sum_{error \in Errors} w_{error} \quad (1)$$

The scale of weights per type of error comes from the WMT21 contest. Two summary tables from (Freitag et al., 2021b) list them exhaustively in Figures 2 and 3.

Severity	Category	Weight
Major	Non-translation all others	25 5
Minor	Fluency/Punctuation all others	0.1 1
Neutral	all	0

Figure 2: en-de and zh-en : Google's MQM error

Severity	Category	Weight
Critical	all	10
Major	all	5
Minor	all	1

Figure 3: en-ru : Unlabel's MQM error

2.3 Visualisations

Figure 4: The majority of translations performed by the systems were assigned the maximum score, namely 0 (no errors identified by a human expert). Only a small proportion of candidate translations multiply errors: for example, for the English-German pairs, 5% of sentences have a score lower than -5. For the Chinese-English and English-Russian pairs, 10% of candidates have a score below -6 and 25% have a score below -2. Figures 5 and 6 shows the distribution of the average score per segment and per domain respectively. Overall, the scores are better for the pair English-German, with a global mean of -0.75, that is coherent with their linguistic proximity. In comparison, the two other pairs English-Russian and Chine-English show an average score of around -1.8. Finally, candidate sentences belonging to the domain of conversation are rated higher than sentences in the other domains, especially the social one.

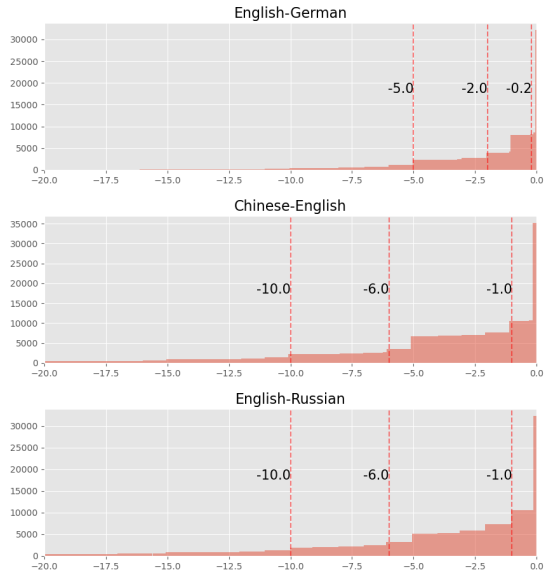

Figure 4: Distribution of the human score for each pair source-target language and quantiles of level 5%, 10% and 25%

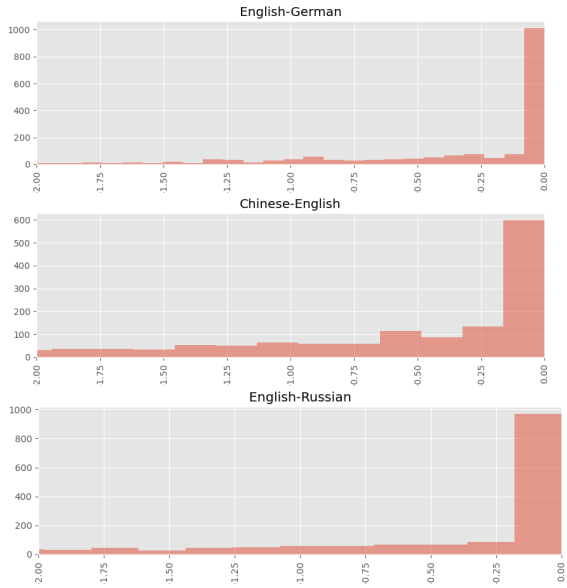

Figure 5: Distribution of the average segment score for each pair source-target language

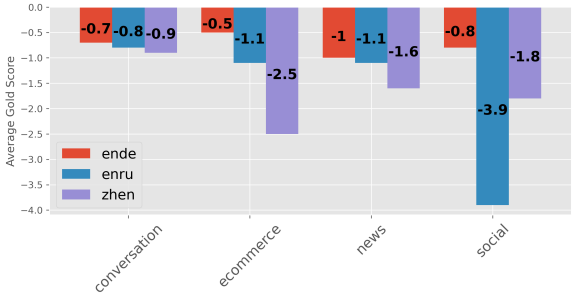

Figure 6: Average score per domain for each pair source-target language

In what follows, we build a benchmark protocol to rank the performance of the existing metrics in MT, considering several criteria all related to correlation with those human scores. The structure of the dataset naturally leads to study this correlation at sentence-level.

3 Experiments Protocol

3.1 Choice of the metrics

The metrics used for our work are gathered in Figure 7. They were choosen among the plethora of metrics which has been developed these last 20 years to address the automatic evaluation of translation, but more generally of text generation tasks commonly identified in NLG, including summarization, data2text generation or story generation.

Historically, the first developed metrics aimed at comparing system and reference translations based on surface forms. One strategy is to study the co-occurrence statistics based on word/character n-gram overlaps and was first introduced by (Papineni et al., 2002). Their measure, BLEU, remains today a dominant metric for MT research as it has been shown to correlate reasonably well with human assessments while staying relatively intuitive, easy to compute and language-independant (Post, 2018). However, BLEU is a geometric mean of unigram to n-gram precisions: any candidate translation without a n-gram match has a per-sentence BLEU score of zero. Therefore, it is a measure that does not work reliably well at sentence level. With methods similar to BLEU, other metrics were then developed. For instance, ROUGE (Lin, 2004), a recall-related measure relying on n-gram overlaps too, initially designed for summarization but adapted for MT tasks through some of its variants, like ROUGE-L or ROUGE-S (Lin and Och, 2004), relying on longest common subsequence and skip-bigram co-occurrence statistics respectively. Other metrics include METEOR (Lavie and Agarwal, 2007), chrF (Popović, 2015), a character n-gram F-score, and its variant chrF+/++ (Popović, 2017) adding short word n-grams (unigrams and bigrams).

Another strategy, alternative to the n-grams, is to compute an edit distance between candidate and reference translations : WER (Nießen et al., 2000) and TER (Snover et al., 2006) are some examples. One of the main weaknesses of all of these text-based metrics is that they do not take into account synonyms and paraphrases. Thus, a candidate having the same meaning of its related reference but with distinct surface forms will be poorly scored.

To tackle this issue, another class of metrics has been developed more recently, comparing system and reference texts based on semantics by relying on words embeddings. These metrics can use static embeddings (not included in the present paper) such as word2vec, or contextualized embeddings (CE), like the BERT-based metrics: BertScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019), DepthScore (Staerman et al., 2021) and BaryScore (Colombo et al., 2021c). Such metrics combine contextualized representations with a distance measure and demonstrated strong generalization capability across tasks.

N-gram based	Edit based	C. Embedding
BLEU, Sacre_BLEU		MOVERScore
ROUGE_1, ROUGE_L	WER	DEPTHScore
ROUGE_S4, METEOR	TER	BARYScore
chrF, chrF_1, chrF++		

Figure 7: Set of the 14 metrics retained

3.2 Choice of the parameters

For each metric, diverse parameters have to be chosen adequatly for allowing legitimate comparisons between them.

- **Preprocessing schemes** User-supplied reference pre-processings like tokenization and normalization schemes have a large effect on scores. For illustration, during many years, BLEU scores between papers could not have been directly compared because of those implicit pre-processings. Quite recently, Sacre_BLEU (Post, 2018) has been designed to settle upon one common BLEU preprocessing-scheme used by the WMT. In

the paper, we consider both the original implementation of BLEU and Sacre_BLEU to look at the eventual gaps in the results. Overall, a basic tokenization has been applied, including a convert to lowercase and for ROUGE metrics, a Porter stemmer procedure that has shown to produce better correlation with adequacy (Lin and Och, 2004).

- **Parameter n for the n-grams** We choose n = 1 for all our n-gram based metrics, to produce more consistent results at sentence level. For chrF, we fix the parameters to their values by default : the character n-gram order to 6 and β to 2. With chrF_1, we modify the value of β to 1 (equal weight put on recall and precision) and with chrF++, a word bigram is added.

- **Pre-trained model** Embedding-based metrics with Neural Networks are obviously dependant on the choice of the pre-trained model. Thus, and as put forward in (Colombo et al., 2021c), it is essential to select one single model in order to produce reliable results; here, we choose the BERT-base-uncased one. Note that fine-tuning BERT representations on datasets like NLI or MultiNLI is possible and would conduct to slightly better results as it was shown in (Zhao et al., 2019; Colombo et al., 2021c).

- **Distance measure** Each of the BERT-based metrics involve a step using a measure of (dis)similarity between the embedded reference and candidate sentences. For both BaryScore and DepthScore, the Wasserstein distance is considered (even if in Staerman et al., 2021 (DepthScore), the AI-IRW depth is considered). For MoverScore, a Word Mover Distance (WMD) is computed between the two sequences of n-gram embeddings of the reference and the candidate. Following the conclusions of (Zhao et al., 2019) finding better results when considering unigrams, we choose n = 1 (same choice as the n-gram metrics)

3.3 Choice of the criteria

In light of the various observations highlighted in the visualization section, we have established a list of criteria to evaluate different metrics:

- Domain Coverage (DC): *Does the metric perform well for translating sentences from various domains (social, e-commerce, conversation, news)?*

- Bad Quality Detection (BQD): *Does the metric effectively detect poor-quality samples (i.e. those with the lowest gold scores)?*

- Segment Level Correlation (SLC): *Does the metric accurately identify the overall quality of a segment (i.e., is its average score per segment well correlated with the gold average score per segment)?*

For each of these three criteria, we restrict the data to a subset of interest: the $\alpha\%$ least well-rated samples for BQD, samples from a specific domain for DC, and all aggregated data at the segment level for SLC. For each criterion, we determine a rank for each metric based on its correlation with the gold score (Pearson, Kendall, Spearman measures).

Considering all of these criteria allows us to obtain a set K of rankings. In order to aggregate these rankings to obtain a final ranking, we rely on (Colombo et al., 2022a) and specifically use the Borda Count technique. Conceptually, this method aggregates ranks as follows:

$$Borda\,Ranks = argsort(\sum_{c=1}^{C} argsort(K_{1,c}, \ldots, K_{N,c}))) \quad (2)$$

where $K_{i,c}$ is the correlation betweeen the ith metric score and gold score for criterion c. Unlike the Kemeny consensus, which proves to be an NP-hard problem, the Borda count is an effective alternative in terms of computational aspects.

4 Results

Figure 12. For the pair of English-German, we obtained the different ranks of the metrics (from 2 to 15) for each of the three criteria. We see that BERT metrics perform reasonably well relatively to the others as they are often ranked among the top 5 for each criterion (materialized by the green area leftside). This result is not surprising as the pre-trained BERT metrics are among the new wave of automatic metrics allowing to handle more robustly synonyms and focusing more and better on the general meaning conveyed by the sentences.

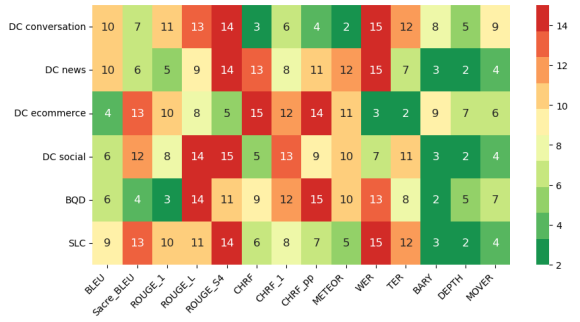

Figure 8: Ranks of each metric along the different criteria – English/German pair

Figure 13. The rankings over the criteria are then aggregated to produce a final ranking of our 14 metrics, here still for the pair English-German. The results for the other pairs are available in the appendix. In all the three couples of sl-tl, the pre-trained BERT metrics outperform all other baselines. Traditional metrics like BLEU, its extension METEOR, and ROUGE_1, show quite good performances relatively to the others, while edit-based metrics appear less suitable, especially WER, historically the older one (2000) and that seems to be outdated today. Consistently with the results of (Colombo et al., 2021c), we found that BaryScore outperforms BertScore for all the three pairs of languages.

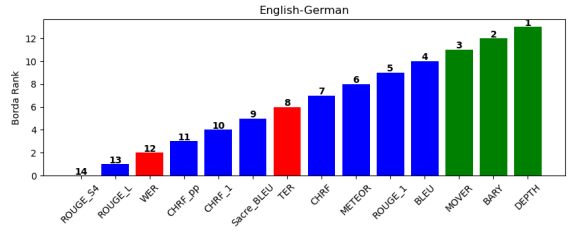

Figure 9: Aggregated Ranks of each metric obtained by Borda's Count procedure – English/German pair

5 Conclusion

The findings of this paper demonstrate the superior performance of the new wave of automatic metrics that rely on contextualized embeddings, compared to traditional metrics, in assessing the quality of Machine Translation output. Specifically, the study revealed that these metrics outperformed traditional metrics in terms of their sentence-level correlation with human scores, as evaluated on the WMT22 dataset.

More generally, these results are know for not only Machine Translation but also other natural language generation tasks (Colombo et al., 2022c; Chhun et al., 2022). The generalization capabilities of contextualized embeddings have been demonstrated across multiple NLG branches, highlighting their potential to improve the evaluation of NLG systems beyond just Machine Translation.

However, there is still much to be explored in the use of these metrics in the context of natural language generation. Future research should aim to investigate the performance of these metrics across a broader range of NLG tasks (*e.g,* conditionnal generation (Colombo et al., 2019, 2021b), multimodal learning (Garcia et al., 2019; Colombo et al., 2021a) and datasets, to further validate their effectiveness in assessing the quality of NLG output. Moreover, there is a need to explore the potential of combining these metrics with other evaluation techniques, such as human evaluation, to provide more comprehensive and accurate assessments of NLG systems.

References

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 605–612.

Chin-Yew Lin. 2004. Rouge: A Package for Automatic Evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231.

Maja Popović. 2015. chrF: character n-gram F-score for Automatic MT Evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. *arXiv preprint arXiv:1804.08771*.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-Driven Dialog Generation. *arXiv preprint arXiv:1904.02793*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. WeinbergerQ, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Alexandre Garcia, Pierre Colombo, Slim Essid, Florence d'Alché Buc, and Chloé Clavel. 2019. From the Token to the Review: A Hierarchical Multimodal approach to Opinion Mining. *arXiv preprint arXiv:1908.11216*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. *arXiv preprint arXiv:1909.02622*.

Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021c. Automatic Text Evaluation through the Lens of Wasserstein Barycenters. *arXiv preprint arXiv:2108.12463*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *arXiv preprint arXiv:2104.14478*.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. 2021a. Improving Multimodal fusion via Mutual Dependency Maximisation. *arXiv preprint arXiv:2109.00922*.

Pierre Colombo, Chloé Clavel, and Pablo Piantanida. 2021b. A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations. *arXiv preprint arXiv:2105.02685*.

Guillaume Staerman, Pavlo Mozharovskyi, Pierre Colombo, Stéphan Clémençon, and Florence d'Alché Buc. 2021. A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions. *arXiv preprint arXiv:2103.12711*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774.

Pierre Colombo, Maxime Peyrard, Nathan Noiry, Robert West, and Pablo Piantanida. 2022b. The Glass Ceiling of Automatic Evaluation in Natural Language Generation. *arXiv preprint arXiv:2208.14585*.

Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022c. Learning Disentangled Textual Representations via Statistical Measures of Similarity. *arXiv preprint arXiv:2205.03589*.

Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M Suchanek. 2022. Of Human Criteria and Automatic Metrics: A Benchmark of the Evaluation of Story Generation. *arXiv preprint arXiv:2208.11646*.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Clémençon. 2022a. What are the best systems? New perspectives on NLP Benchmarking. *arXiv preprint arXiv:2202.03799*.

A Results for the Chinese-English pair

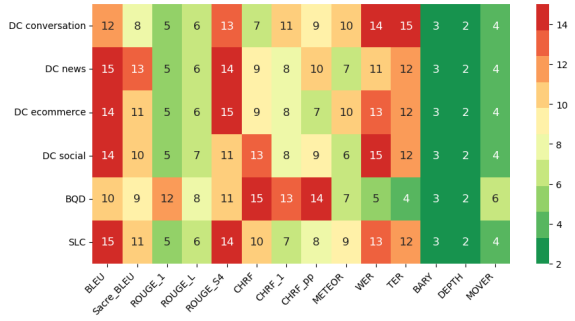

Figure 10: Ranks of each metric along the different criteria – Chinese/English pair

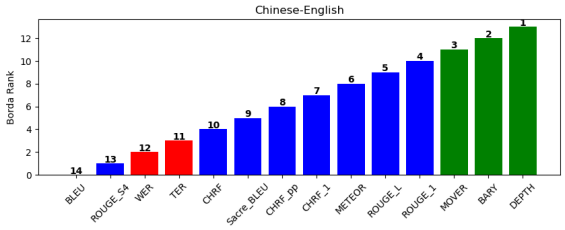

Figure 11: Aggregated Ranks of each metric obtained by Borda's Count procedure – Chinese/English pair

B Results for the English-Russian pair

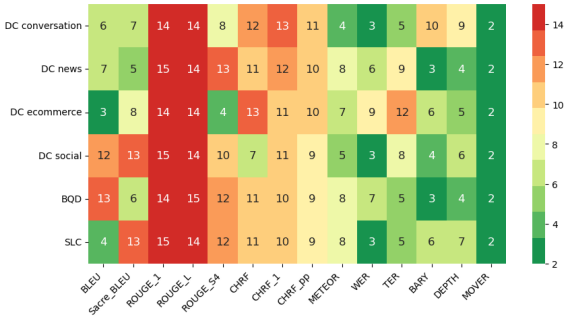

Figure 12: Ranks of each metric along the different criteria – English/Russian pair

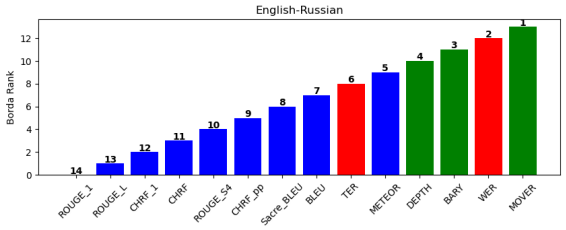

Figure 13: Aggregated Ranks of each metric obtained by Borda's Count procedure – English/Russian pair